PROJET 7

Implémentez un modèle de scoring

PLAN

- Présentation du projet
- Analyse exploratoire des données
- Pré-traitement des données
- Approche de modélisation
- Présentation du dashboard
- Conclusion

Présentation du projet

- L'entreprise « Prêt à dépenser » propose des crédits à la consommation pour des personnes ayant peu ou pas du tout d'historique de prêt.

- <u>Objectifs :</u>

- Développer un modèle de scoring de la probabilité de défaut de paiement du client
- Développer un dashboard interactif

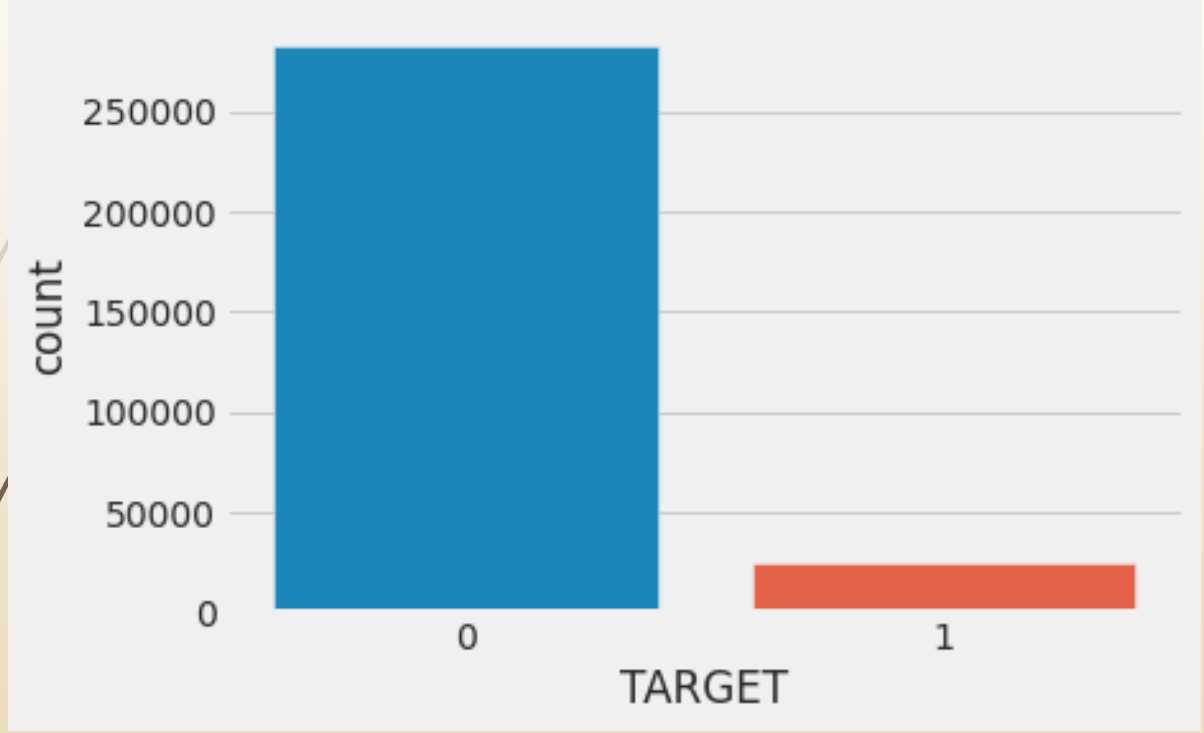
Analyse exploratoire des données

Prêt à dépenser

- 8 jeux de données contenant des informations concernant 307 511 clients : âge, situation familiale, lieu de résidence, emploi…

- TARGET
  - 0 si le prêt est remboursé à temps
  - 1 si le prêt n'est pas remboursé à temps

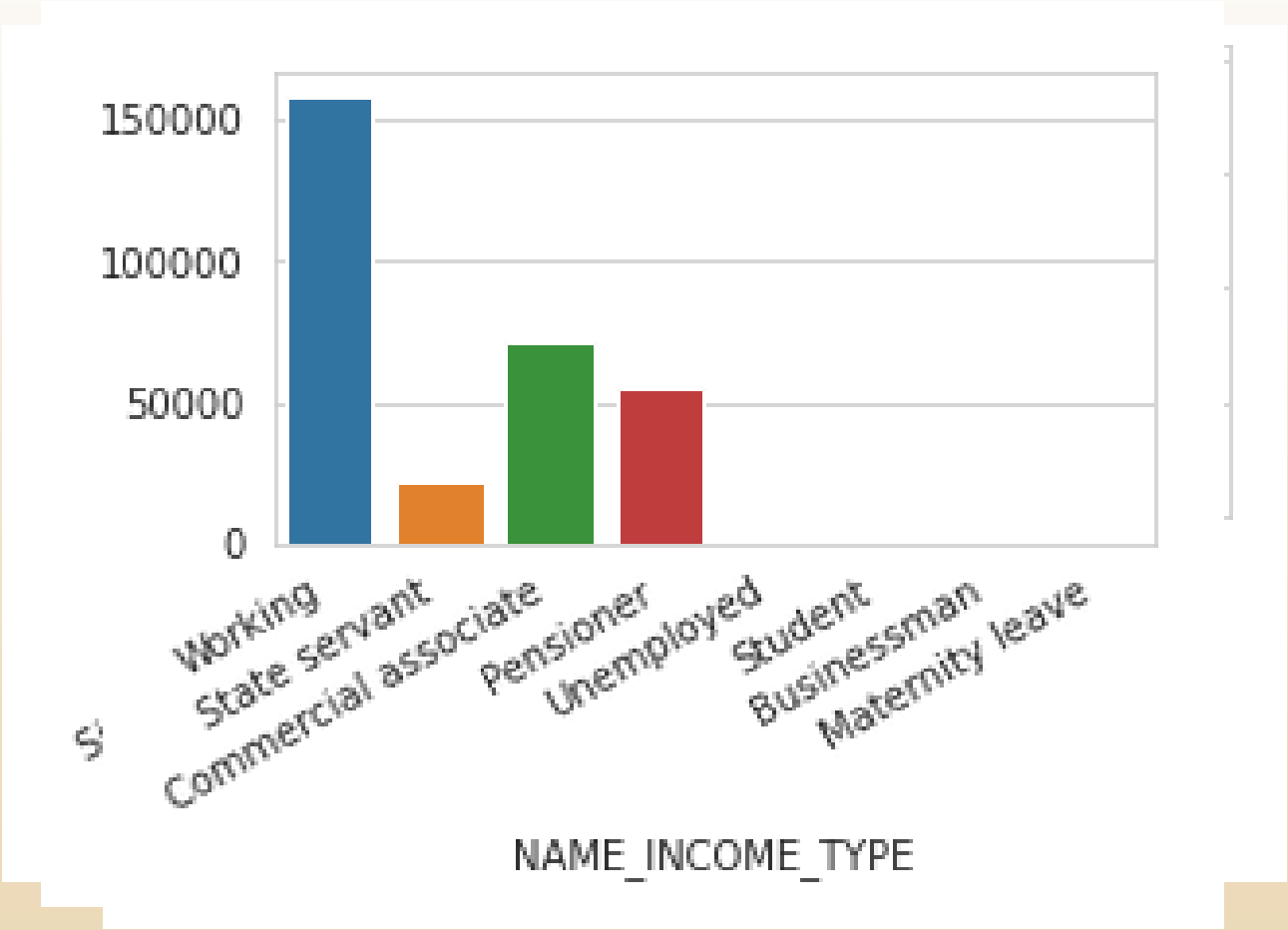
Analyse exploratoire des données



La distribution de la variable cible montre que le nombre de prêts classés « remboursés à temps » est nettement supérieur à ceux qui ne le sont pas. Il faudra remédier à ce déséquilibre des classes.

Analyse exploratoire des données

Distributions des variables catégorielles



Nous avons 16 variables catégorielles dont :

- Le type d'emploi ;
- Le statut familial ;
- Le niveau d'étude

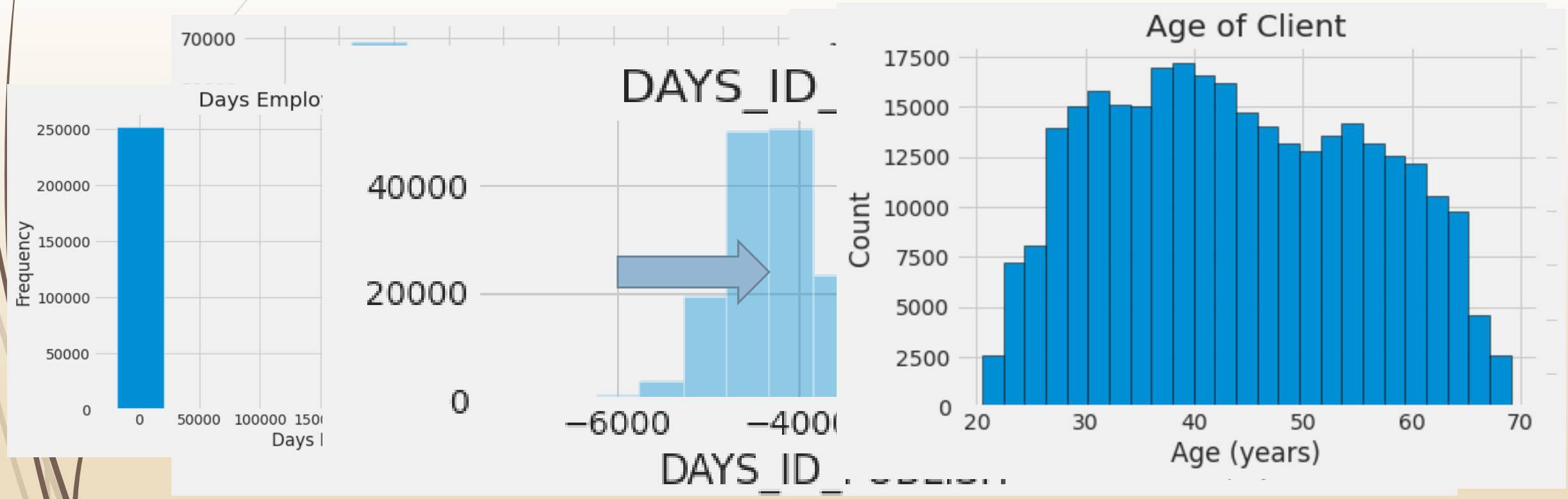
Analyse exploratoire des données

Distributions des variables numériques

Prêt à dépenser

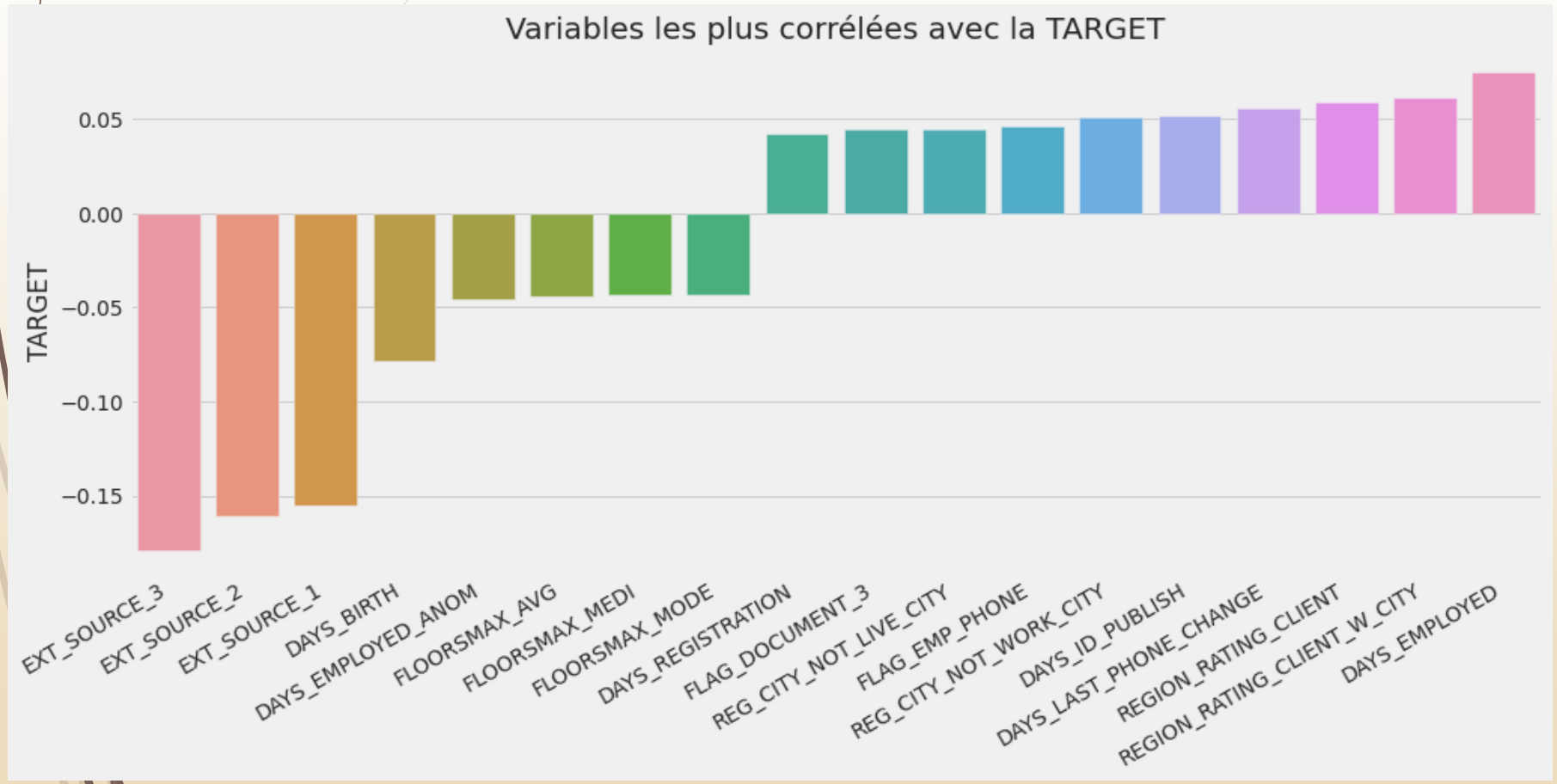
Le nombre de jours depuis lequel le client est employé :

Le genre de la carte d'identité :



La variable « Days_employment » présentait 55 374 données égales à 365 243 jours, ce qui représentent environ 1 000 ans. Nous avons remplacé ces valeurs aberrantes par des valeurs manquantes.

Analyse exploratoire des données

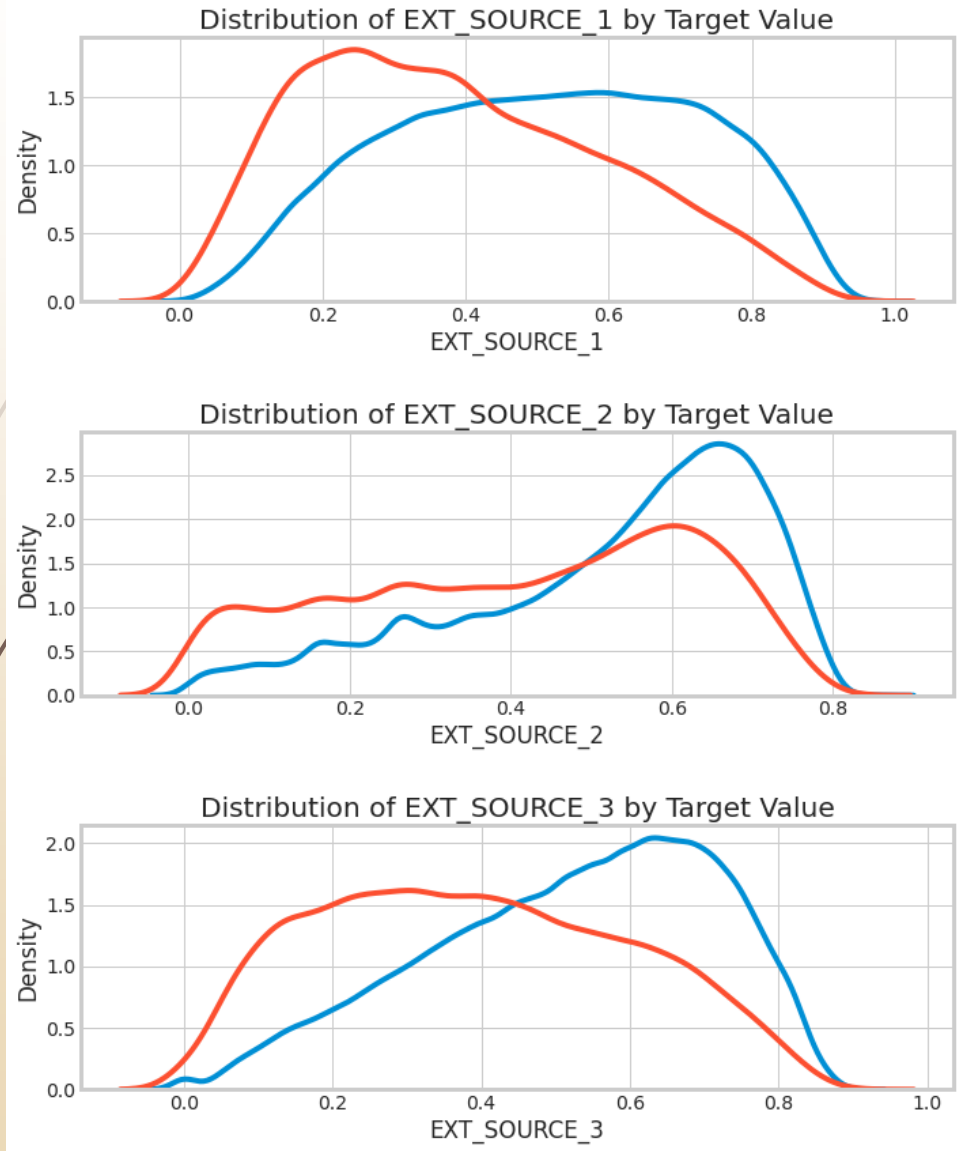


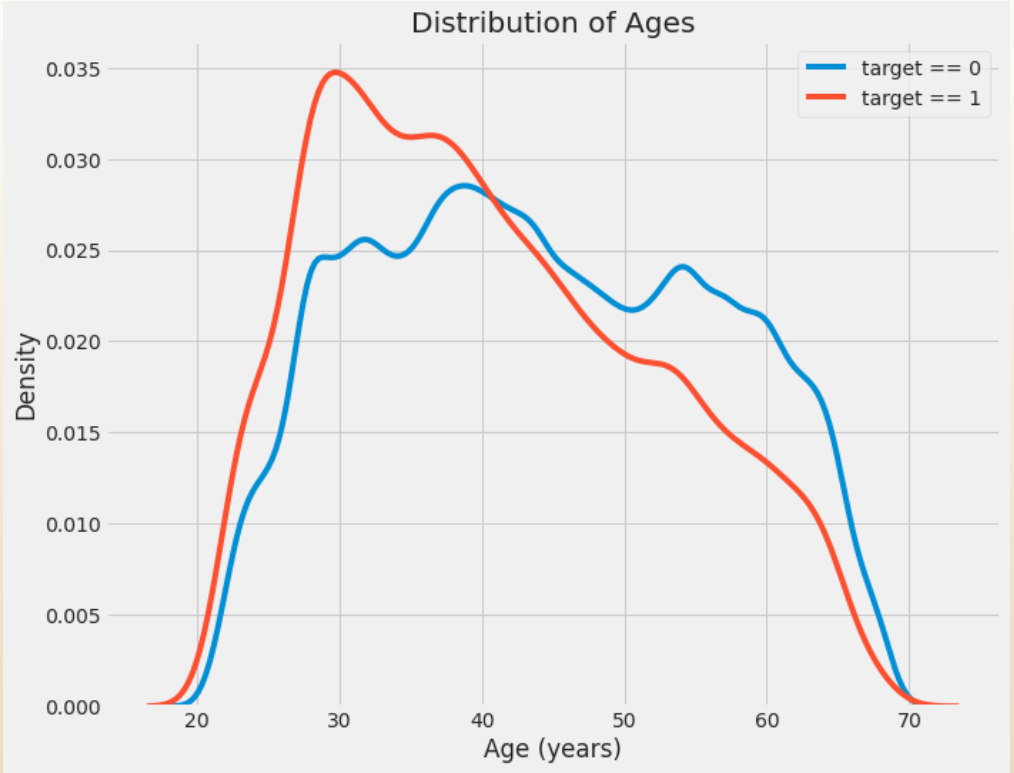
Les variables « EXT_SOURCE » et « DAYS_BIRTH » sont les variables les plus fortement anti-corrélées avec notre variable cible.

La variable « DAYS_EMPLOYED » est la variable la plus fortement corrélée avec notre variable cible.

Analyse exploratoire des données

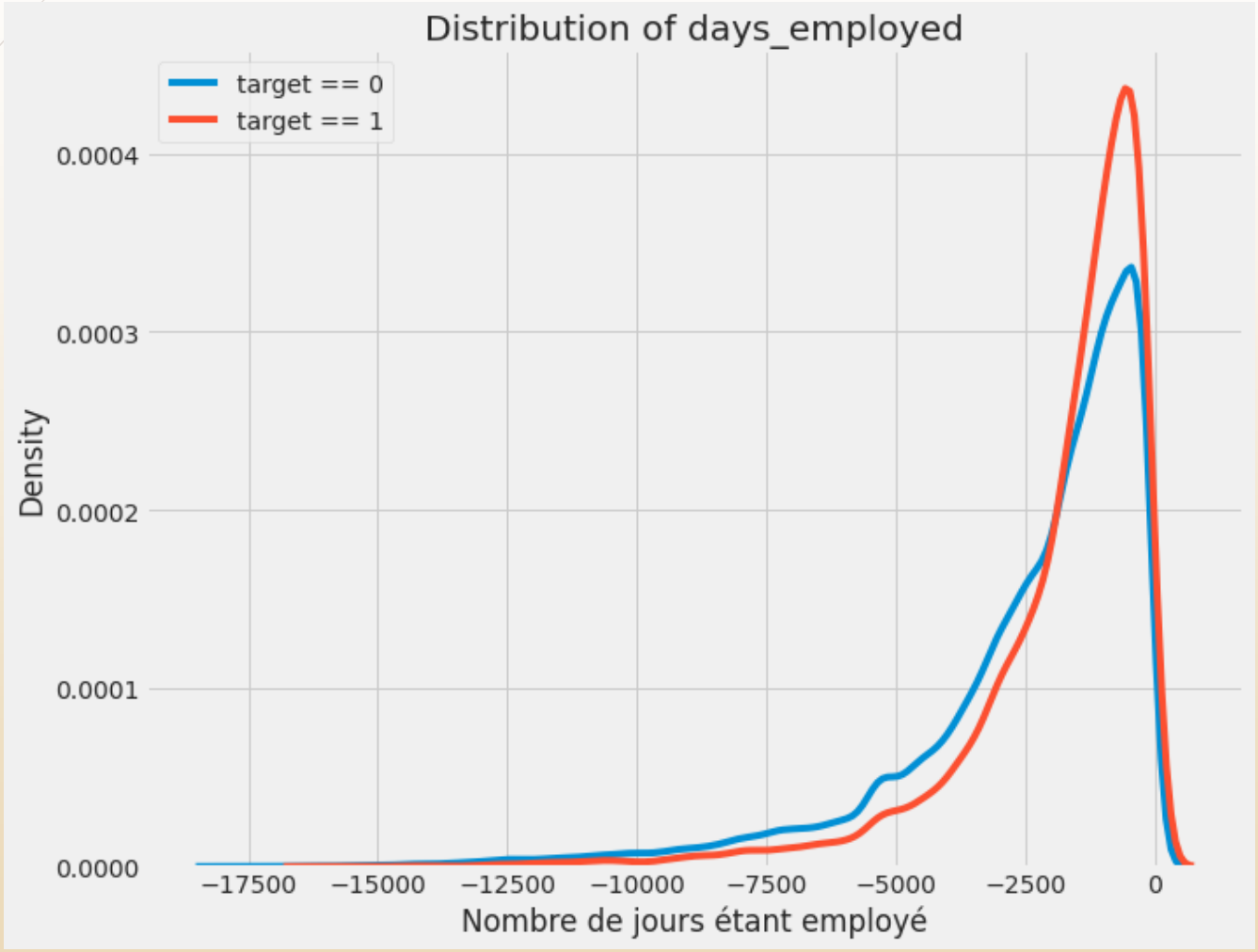
Prêt à dépenser





Variables anti-corrélées avec la variable cible.

Analyse exploratoire des données



Variable corrélée avec la variable cible.

Pré-traitement des données

Séparation du jeu de données

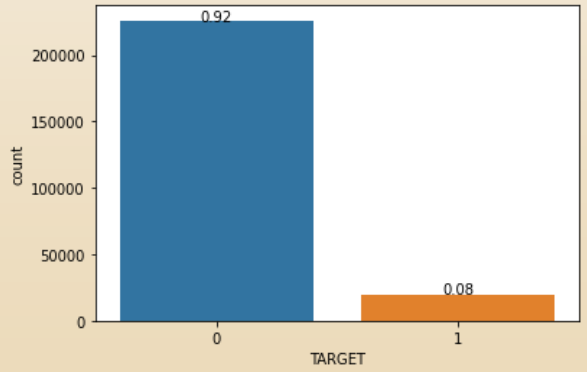
Prêt à dépenser

Jeu de données contenant 307 511 individus

80 %

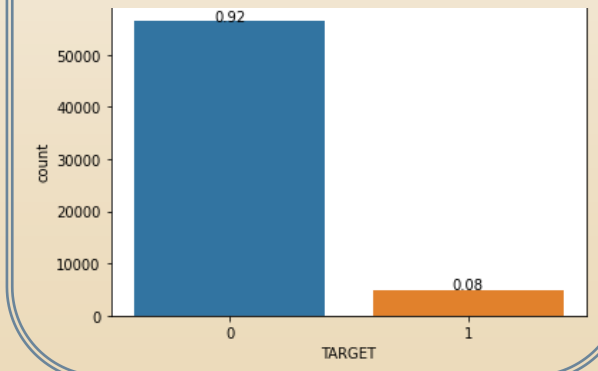
20 %

Jeu d'entrainement contenant 246 008 individus



Proportion des classes 0 et 1 conservées

Jeu de test contenant 61 503 individus



Pré-traitement des données

Pipeline

Prêt à dépenser

ColumnTransformer

Variables catégorielles

- SimpleImputer()
- OneHotEncoder()

Variables numériques

- SimpleImputer()
- Normalisation : StandardScaler() ; MinMaxScaler() ; MaxAbsScaler().

Approche de modélisation

Pipeline

Prêt à dépenser

ColumnTransformer

Variables catégorielles

- SimpleImputer()
- OneHotEncoder()

Variables numériques

- SimpleImputer()
- Normalisation : StandardScaler() ; MinMaxScaler() ; MaxAbsScaler().

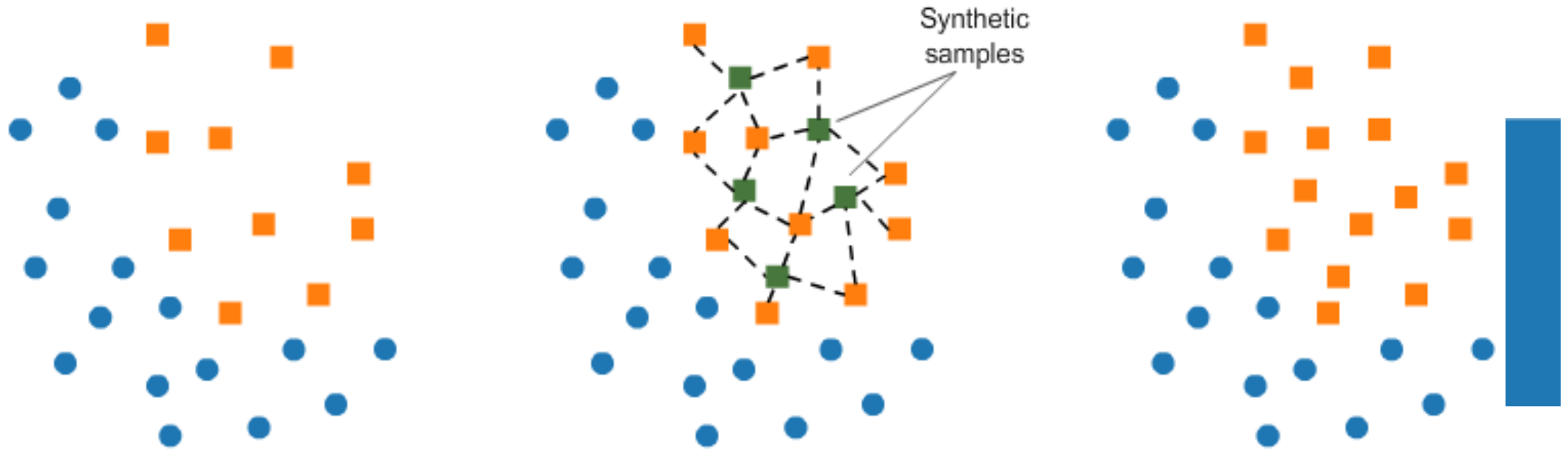
- Sampling : RandomOverSampling ; RandomUnderSampling ; SMOTE
- Feature selection : SelectKBest()
- Modèle

Approche de modélisation

Pipeline

Prêt à dépenser

RandomUnderSampling | RandomOverSampling | SMOTE



Approche de modélisation

Choix de la métrique

Prêt à dépenser

Matrice de confusion

		Prédiction	
		0 (sans défaut)	1 (en défaut)
Réalité	0 (sans défaut)	Vrais négatifs	Faux positifs
	1 (en défaut)	Faux négatifs	Vrais positifs

Pertes réelles pour la banque puisque le crédit client est accepté mais ne sera pas remboursé à temps.

Nous souhaitons minimiser les taux de faux négatifs et de faux positifs.

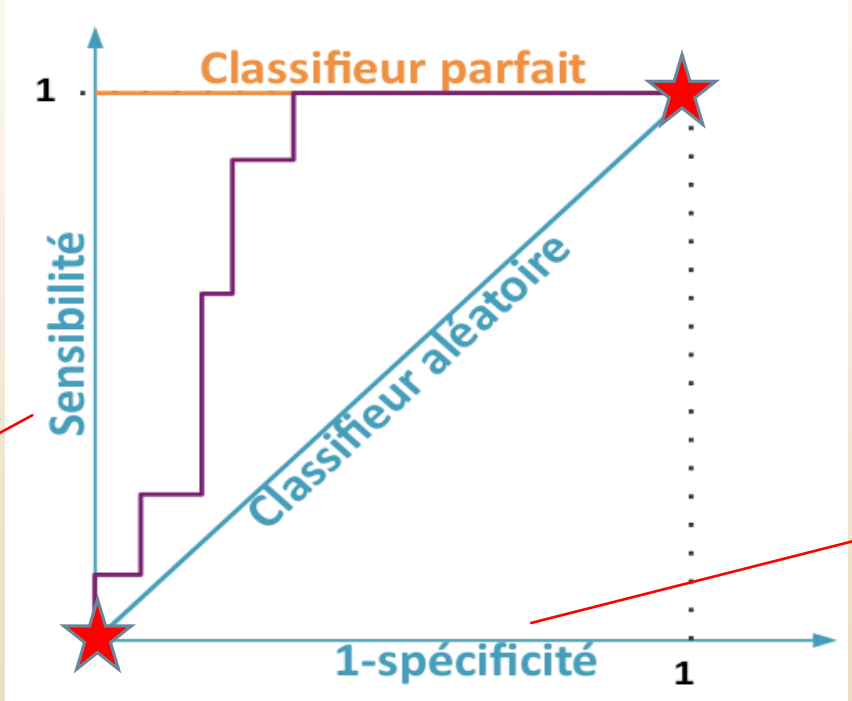
Pertes d'opportunités puisque le crédit client est refusé alors qu'il aurait été en mesure d'être remboursé.

Approche de modélisation

Choix de la métrique

Prêt à dépenser

La courbe ROC (Receiver Operating Characteristic)



On cherche à maximiser le score AUROC (Area Under the Curve) qui correspond à l'aire sous la courbe ROC.

$Sensibilité = \frac{VP}{VP+FN}$ c'est le taux de vrais positifs (positifs bien identifiés),

$1 - Spécificité = \frac{FP}{FP+VN}$
c'est le taux de faux positifs

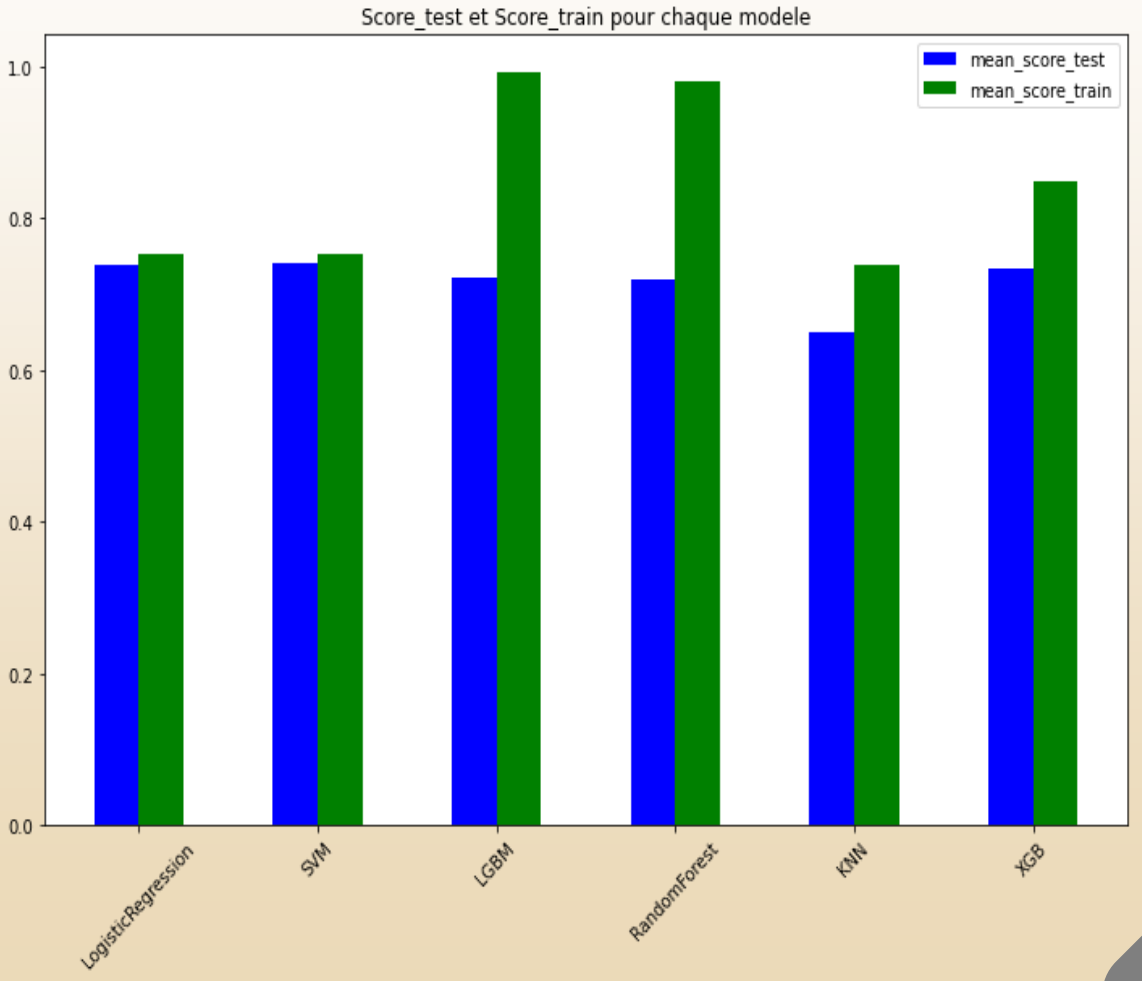
★ Seuil fixé à 0, tous les prêts seront classés négatifs (non remboursés) donc aucun vrai positifs et aucun faux positifs.

Approche de modélisation

Choix du modèle

Prêt à dépenser

	model	mean_score_train	mean_score_test	mean_time_train	mean_time_test
0	gridlogistic	0.752420	0.739591	0.972953	0.067023
1	gridsvc	0.751906	0.740065	2.017171	0.068047
2	gridlgbm	0.992485	0.721089	3.116567	0.087681
3	gridrandomforest	1.000000	0.722695	3.993857	0.131265
4	gridknn	0.739448	0.650790	0.536391	0.167575
5	gridxgb	0.848218	0.733768	3.041840	0.074816



Nous choisissons le modèle **Logistic Regression Classifier.**

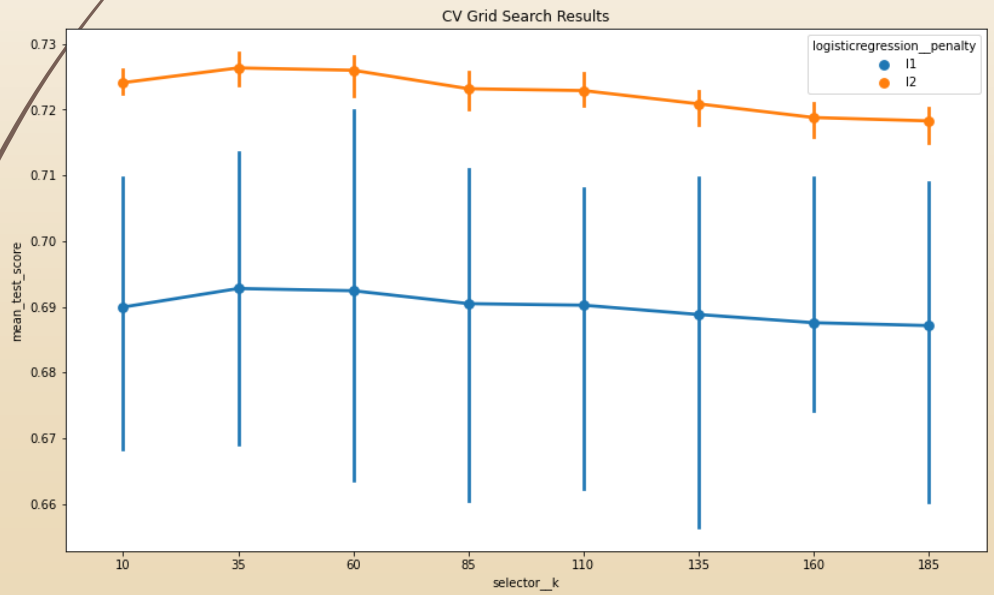
Approche de modélisation

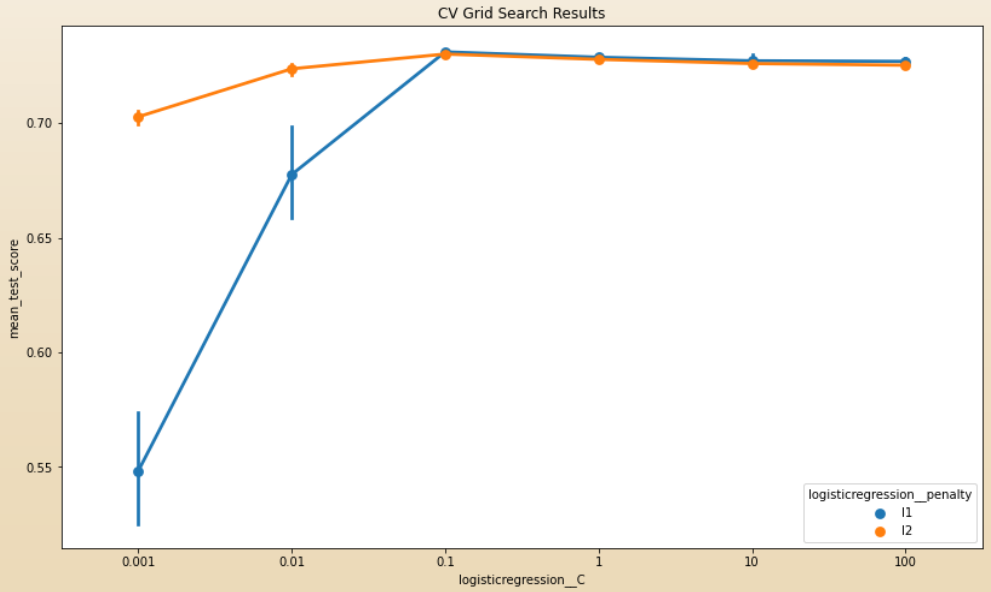
Optimisation de Logistic Regression Classifier

Prêt à dépenser

Grille de paramètres:

```
param_logistic = {'preprocess__pipeline1__scaler': [StandardScaler(), MinMaxScaler(), MaxAbsScaler()],
                  'sampler': [rus, ros, smote],
                  'selector__k': list(range(10,200,25)),
                  'logisticregression__C':[0.001, 0.01, 0.1, 1, 10, 100],
                  'logisticregression__penalty' : ['l1', 'l2'],
                  'logisticregression__solver':['newton-cg', 'lbfgs', 'saga']}
```



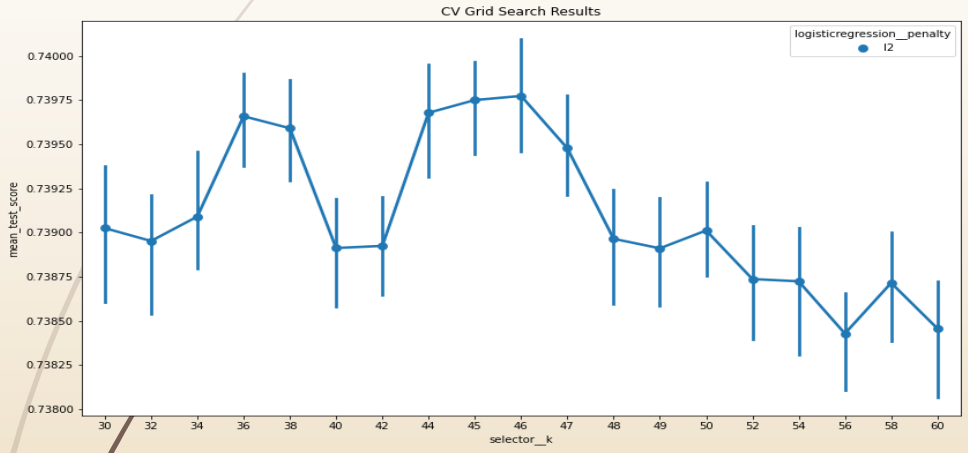


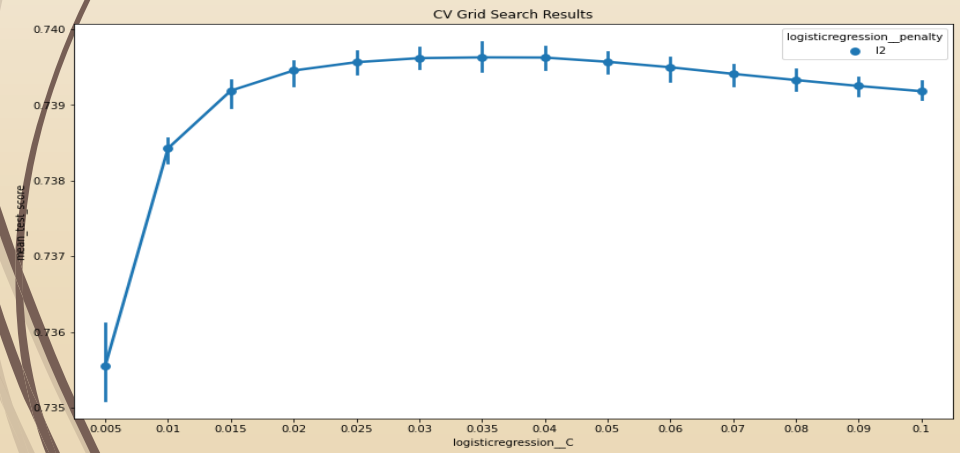
Approche de modélisation

Optimisation de Logistic Regression Classifier

Prêt à dépenser

Choix des valeurs pour k (nombre de features pour SelectKBest) et C (Logistic Regression Classifier) :





Grille de paramètres présentant le meilleur résultat:

```
{'logisticregression__C': 0.03,
 'logisticregression__penalty': 'l2',
 'logisticregression__solver': 'newton-cg',
 'preprocess__pipeline-1__scaler': MaxAbsScaler(),
 'sampler': RandomOverSampler(random_state=0),
 'selector__k': 46}
```

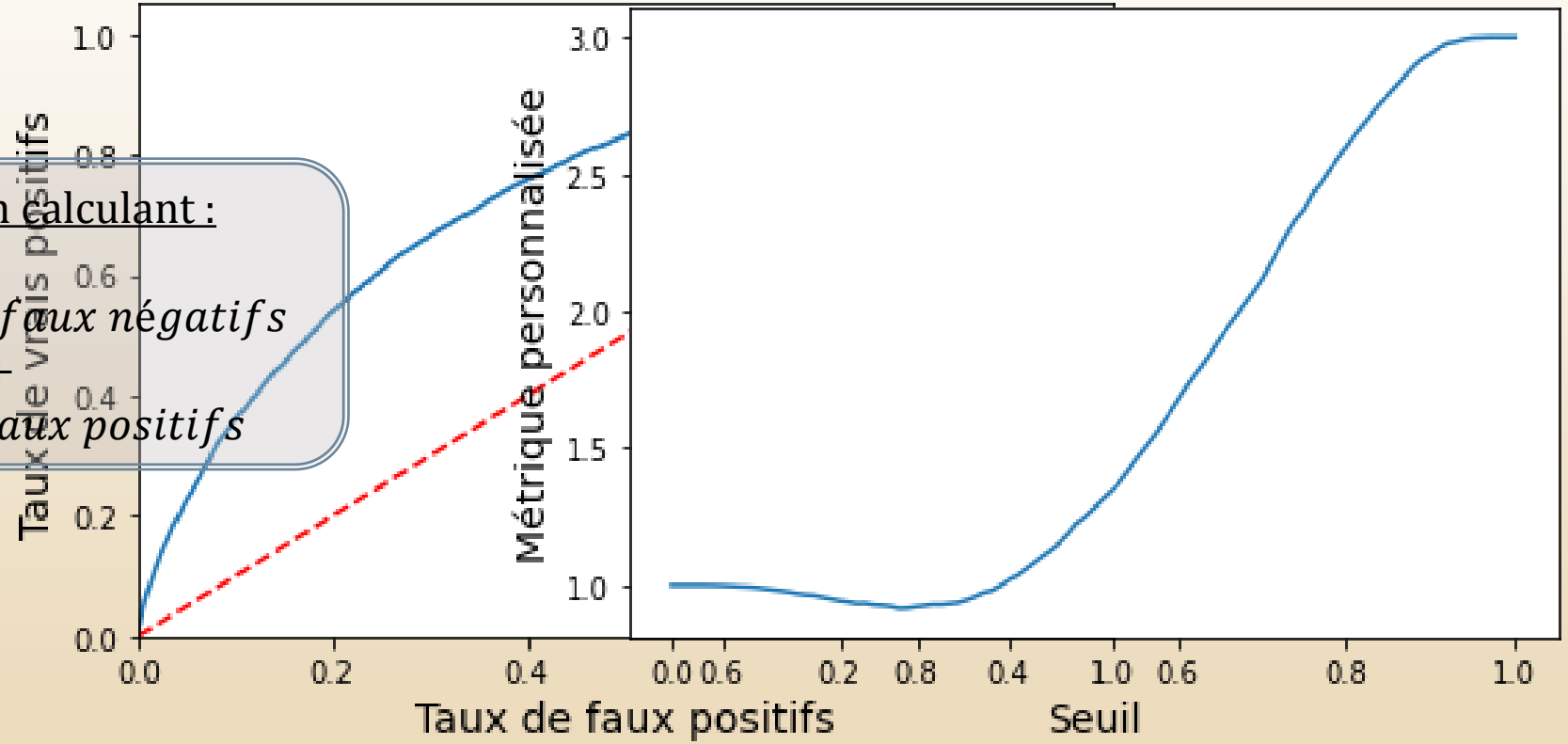
Nous obtenons un score AUROC d'environ 0.741 sur le jeu de validation.

Approche de modélisation

Choix du seuil

Prêt à dépenser

Courbe ROC obtenue avec le modèle Logistic Regression



Choix du seuil de décision

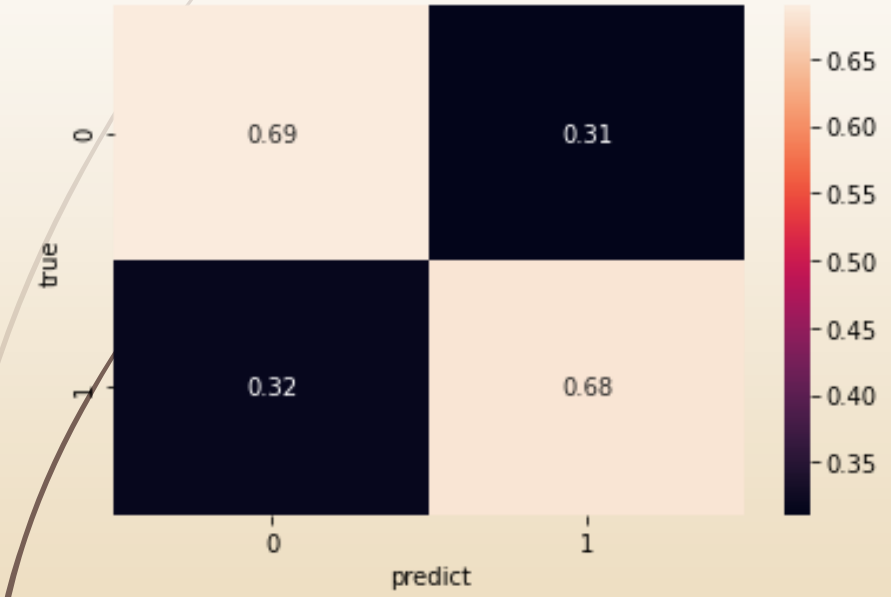
Choix du seuil en calculant :

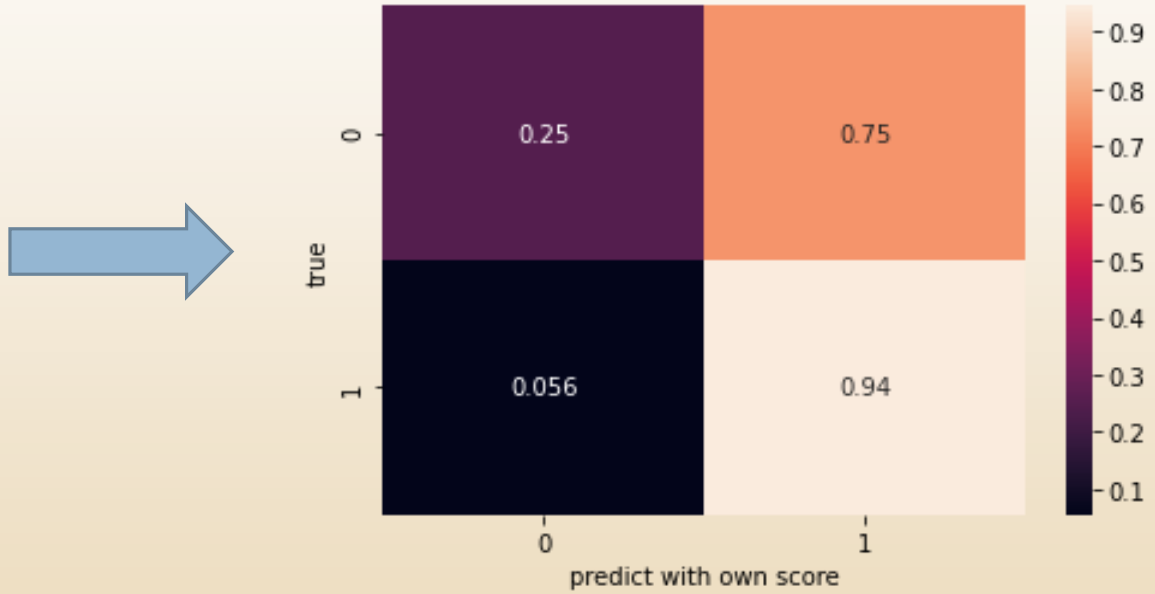
$$3 \times le\ taux\ de\ faux\ négatifs + le\ taux\ de\ faux\ positifs$$

Nous choisirons un seuil égal à 0.27.

Approche de modélisation

Impact sur la matrice de confusion





Sans la métrique personnalisée, seuil à 0.5.

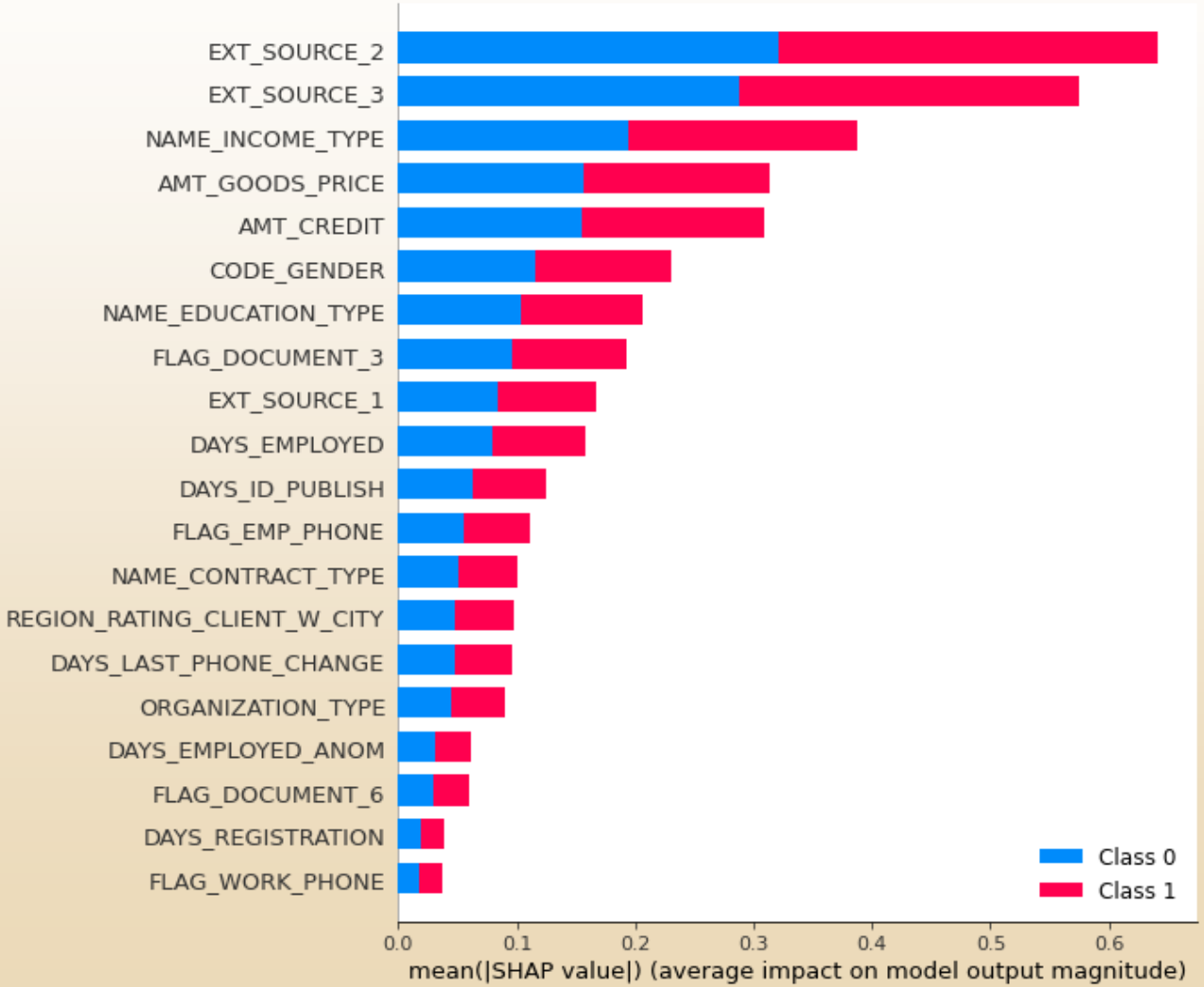
Avec la métrique personnalisée, seuil à 0.27.

Il y a beaucoup moins de faux négatifs.

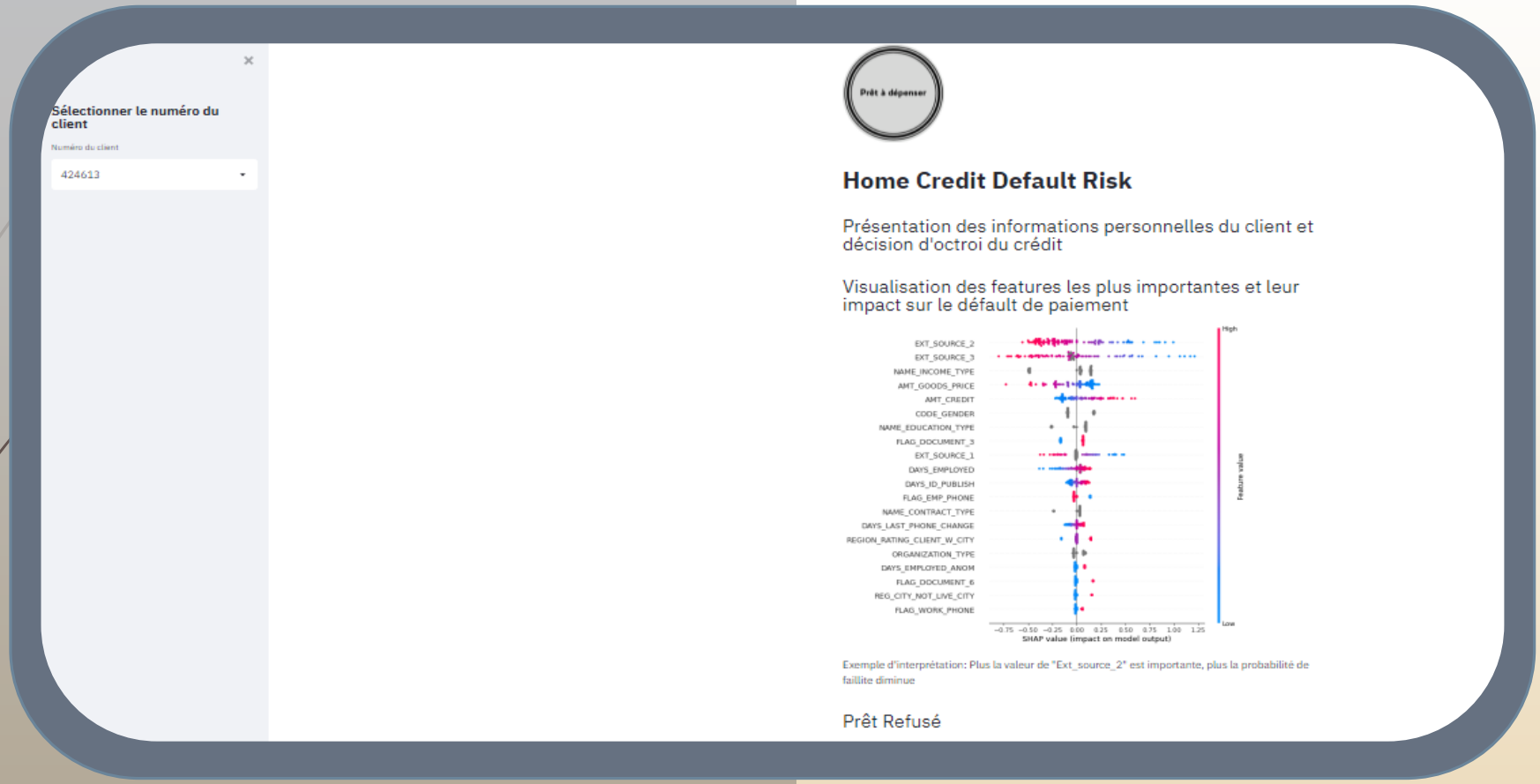
Interprétabilité du modèle

Prêt à dépenser

Variables les plus importantes pour les prédictions de notre modèle :



Présentation du dashboard



[Lien vers le dashboard : https://home-credit-sei.herokuapp.com/](https://home-credit-sei.herokuapp.com/)

CONCLUSION

- Utiliser davantage de données
- Compréhension des features
- Amélioration du dashboard